Methods 89 (2015) 54–63

Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Review Article

XL-MS: Protein cross-linking coupled with mass spectrometry

Andrew N. Holding

Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Robinson Way, Cambridge CB2 0RE, UK

ARTICLE INFO

Article history:
Received 30 January 2015
Accepted 8 June 2015
Available online 12 June 2015

Keywords:
Proteomics
Mass spectrometry
Structure
Quantitation
Systems biology
Cross-linking
XL-MS

ABSTRACT

With the continuing trend to study larger and more complex systems, the application of protein cross-linking coupled with mass spectrometry (XL-MS) provides a varied toolkit perfectly suited to achieve these goals. By freezing the transient interactions through the formation of covalent bonds, XL-MS provides a vital insight into both the structure and organization of proteins in a wide variety of conditions. This review covers some of the established methods that underpin the field alongside the more recent developments that hold promise to further realize its potential in new directions.

1. Introduction

1.1. Mass spectrometry and protein interactions

The application of mass spectrometry to the elucidation of protein interactions has led to the development of a wide range of methods. These include detection of non-covalent interactions in the gas-phase [1], the identification of protein tertiary structure using reactive species [2,3], hydrogen–deuterium exchange [4,5] for conformational studies, and the use of cross-linking reagents for the elucidation of both tertiary and quaternary structure and identifying protein–protein interactions. As these techniques all provide different information about the system being studied, it is further possible to then combine these technologies [6]. The popularity of these methods has only grown with the increasing desire to focus on larger and more complex systems which typically are not amenable to established atomic resolution studies [7]. Further benefits of these techniques are: the ability to apply them at, or near to, physiological conditions; they require significantly less material; and they are less stringent on the purity than NMR or X-ray diffraction.

The success of elucidating protein structure with XL-MS has led to a large number of software solutions and methods focused on this application (reviewed by Sinz [8,9]; Rappsilber [10]; Sharon [11]; and protocols by Schmidt et al. [12] and Leitner et al. [13]). The overall workflow of cross-linking combined with mass spectrometry can, however, be applied more broadly. With an understanding of the basic principles, novel data can often be obtained without a significant investment in developing the methodology. This review covers this process to provide an introduction to the methods available and the challenges involved, alongside previously unpublished data on two alternative cross-linking strategies.

1.2. Cross-linking in the elucidation of protein interactions

Some of the earliest reports of applying bi-functional reagents to the study of protein structure focus on how they alter the physical properties of protein fibers [14]. By the 1960s, the same regents were being used to probe the membrane proteins on the surface of cells, but the interpretation of the data was very limited [15]. It was the combination of these reagents with other analytical techniques, including SDS–PAGE [16] and enzymatic digestion [17,18], during the 1970s that showed their potential to analyze quaternary structure and protein interactions.

Today, protein cross-linking is commonly undertaken using homo-bi-functional NHS-esters [19], which form bridges between lysine residues that are in close proximity to each other. The range of this interaction is often controlled by varying the length of the space between the two functional groups. Once formed, the covalent bridges between proteins provide a route to analyze otherwise transient interactions.

1.3. In vivo cross-linking

This process of stabilizing otherwise undetectable interactions has been successfully applied *in vivo*. The use of membrane

E-mail address: andrew.holding@cruk.cam.ac.uk

http://dx.doi.org/10.1016/j.ymeth.2015.06.010
1046-2023/

permeable reagents allows the reagents to permeate the cells with little disruption to active protein organization. To achieve this, cells containing the complex of interest are washed and then incubated with the appropriate cross-linking reagent. The cross-linked reaction must then be quenched and the complex of interest can be purified by immunoprecipitation before analysis by immunoblotting or similar analytical techniques to identify interacting proteins [20–22].

1.4. Mass spectrometry in combination with protein cross-linking

As demonstrated, the elucidation of binding interactions at a protein level continues to be a clear utility of cross-linking reagents. This, combined with ongoing advancement in mass spectrometry over the last decade, is still providing significant opportunities to develop these processes; in particular, increasing sensitivity and accuracy while at the same time notably reducing the acquisition time required to obtain mass spectra [23,24].

It is these advances that form the basis of many of the new methods being developed to apply protein cross-linking to find novel co-factors within specific complexes [25] or even targeted to a specific genomic locus [26,27].

1.5. Structural characterization of proteins and protein complexes

While the application of cross-linking combined with mass spectrometry to provide protein level information continues to be of great value, more detailed analysis is possible.

In a standard shotgun proteomic analysis, the protein sample is digested in peptides by a specific protease, often trypsin. The peptides are then separated by liquid chromatography directly interfaced with a mass spectrometer. In each cycle, first a survey scan is conducted to identify ions likely to be peptides. Typically, the use of data-dependent acquisition would then allow for these ions to automatically undergo some form of fragmentation to provide sequence information. The masses and fragmentation patterns detected by the mass spectrometer are then submitted to a database search against either a general or a species specific list of known proteins [28–31].

Identification is established by the presence of peptides with a sequence that maps on to the parent protein sequence. To provide confidence in the result, a reasonable cut-off is typically set for the minimum number of unique peptides detected that correlate to a particular protein. Further confidence can be gained by using a decoy database to establish a false discovery rate at either the peptide or protein level (reviewed by Choi and Nesvizhskii [32]).

It has been possible to extend and develop these methods by including the possibility for the identification of cross-linked peptides (see Fig. 1). The characterization of these peptides provides the opportunity to gain high-throughput structural insights of protein structure and organization [33]. The difficulties arise, however, with the cross-linked peptides not being identifiable with standard shotgun proteomic software as the fragmentation patterns differ from traditional peptides. Further, each interaction will only be represented by a small number of cross-linked peptides compared to the number of linear peptides as the cross-linking process is nonstoichiometric, making detection of these species difficult. Despite the promises of detailed structural data, the limited number of peptides identified has resulted in the technology not yet living up to expectations. Furthermore, the idea of varying the length of cross-linking reagent space length to provide a molecular ruler has been found to give inconsistent correlation to the range of the interaction detected. Nevertheless, the information gained about structure of targets that have not been amenable to other strategies can be of real benefit to research.

1.6. Identification of cross-linked peptides

Within a cross-linked proteomic sample, there are a combination of different species generated (see Fig. 2). The majority of these will be linear peptides, but typically several cross-linking products will also be formed. The generation of cross-linked peptides can be either inter- or intra-protein, both of which can provide useful information to aid structural analysis. In our study of the architechuture of the Pol III–Clamp–Exonuclease complex [34], we used the identification of intra-protein cross-links to validate against the known atomic structures of each of the subunits and to provide confidence in the inter-protein cross-links characterized. Alongside the formation of cross-linked peptides are the generation of loop-links, which occur when no proteolytic site exists between two cross-linked residues or because the proteolytic site is not accessible. Generally, the information provided by these loop-link sites is less useful than that provided by inter- or intra-protein cross-links and the choice on the cross-linking reagent used will impact the relevance of the information. Finally we see the production of monolinks, which are produced by the

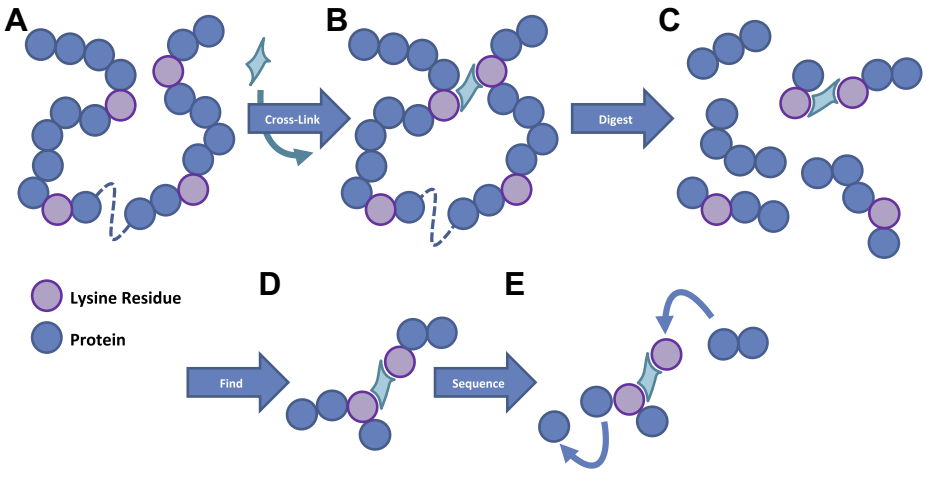

Fig. 1. The typical analysis of a cross-linked sample by shotgun proteomics. (A) The protein or proteins are incubated with a residue specific cross-linking reagent. (B) Residues within the range of the cross-linking reagent are then covalently bonded and transient interactions are stabilized. (C) The protein is then digested by a specific protease to form peptides. (D) Data-dependent acquisition is used to identify peptides as they elute from an HPLC directly coupled to the mass spectrometer. (E) The identified peptides are then fragmented to provide sequence specific information.

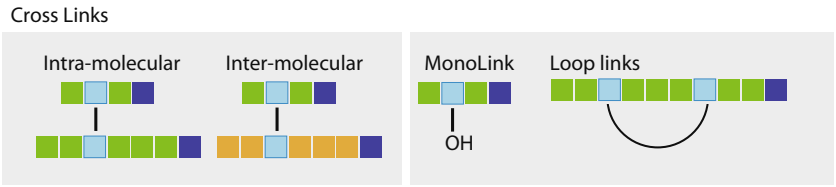

Fig. 2. Schematic diagram of the products from the proteolysis of cross-linked protein samples. Squares represent amino acids. Green squares are residues from the target protein, orange squares are from an interactor, and blue squares represent a proteolytic cleavage site. A black line represents the cross-linking moiety.

hydrolysis (or sometimes amination) of the cross-linking reagent to produce modified peptides; depending on the cross-linking reagent used, these can provide an insight into the location of surface and solvent accessible residues.

The level of complexity within the sample is further confounded by the wide variety of possible inter- and intra-protein cross-links that can form within a protein complex. The digestion of a standard proteomic sample typically results in a small amount of variation in the digestion products between peptide molecules due to so-called 'missed cleavages' (see Fig. 3) and the potential for post-transnational modification.

When building a database from a known proteome to undertake a proteomic database search, it is typical to include two or more missed cleavages within the search parameters. The impact of this is not computationally intensive, yet if we apply a similar model to the analysis of samples containing cross-linked proteins, it demonstrates one of the major challenges in this field. To undertake a database search, it should include all potential cross-linked peptides. The complexity of this task is demonstrated by trying to model these potential products. It should be apparent that every residue containing a side chain that is reactive to the cross-linking reagent can potentially form a covalent bond to every other reactive amino acid side chain if we have no prior knowledge of the structure. This alters the search space from being a linear problem (e.g. n proteins = $k \times n$ peptides) to an algorithm that scales much more poorly as n proteins gives rise to a solution essentially proportional to n^2 cross-linked peptides [35]. This means the inclusion of complicating factors like missed-cleavages or post-translational modifications results in a substantial increase in the number of potential cross-linked products. The issue of complexity is further exacerbated as most common cross-linking reagents react specifically with the basic side chain of lysine residues. The process of cross-linking therefore actively blocks the digestion of trypsin at these sites.

These challenges are not insurmountable, but it has led to the development of many novel strategies and methods. These take the form of three main strategies: to enrich for the peptides of interest over linear peptides; to develop new cross-linking reagents; and for both methods and technologies to aid in the detection of the cross-linked peptides from complex mixtures.

2. Methods

2.1. Peptide-based enrichment

2.1.1. Variation of proteolytic enzyme

Trypsin is used almost ubiquitously in proteomics [36] due to its high specificity and the convenience of the peptides it generates. As the majority of tryptic peptides contain either an arginine or lysine at the C-terminus, this aids peptide ionization efficiency and detection of product ions. The disadvantage of this is the majority of cross-linking reagents are lysine reactive and therefore reduce digestion efficiency. To address this, an evaluation of alternative proteases—including Lys-C, Lys-N, Glu-C and Aps-N—on

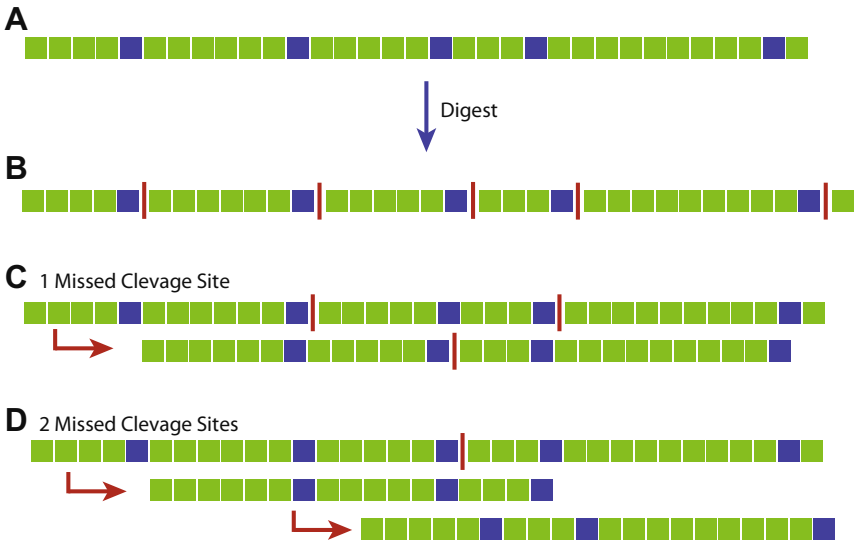

Fig. 3. Schematic representation of a protein undergoing proteolysis and illustrating the complexity of products arising from 'missed cleavage sites'. (A) Illustrates a full-length protein before enzymatic digestion. The simplest solution is a protease that will cleave at all points in the protein sequence that match the specificity of the enzyme – shown in blue. In the case of trypsin, these would be any lysine or arginine residues that are not followed by a proline residue (B). In the majority of samples, however, analysis of the products shows the presence of partial proteolysis products; these are the result of 'missed cleavages' (C). Care has to be taken when modeling these peptides, as for a single missed cleavage there are two series of these peptides. One originates sequentially from the start of the protein from cutting at alternate cleavage sites, and the second is generated in the same manner from the second cleavage site. A similar method is used to generate peptides from the result of two missed cleavage sites (D), generating three series of potential products.

protein samples treated with disuccinimidyl suberate was undertaken by Leitner et al. [37]. The work showed that a significant increase in the number of cross-links could be detected by changing the enzyme used in sample preparation.

2.1.2. Size exclusion chromatography

A second part of the study by Leitner et al. [37] investigated the application of size exclusion chromatography to aid the purification of cross-link peptides within samples. Analysis of the eluted peptides showed that the cross-linked peptides could be detected at higher concentrations in the earlier fractions and therefore size exclusion chromatography promoted their enrichment over modified or unmodified peptides. For tryptic digests, this led to an increase from 44 to 150 detectable cross-linked peptides when the processed and unprocessed samples were compared.

2.1.3. Strong cation exchange

An alternative to the use of size exclusion has been the development and optimization of strong cation exchange (SCX) based enrichment strategies [7,38]. Unlike size exclusion chromatography, the fractionated products of the process contain high salt content and it is necessary to undertake further purification before analysis by mass spectrometry, although some of this can be mitigated by the use of volatile salts. The process relies on the cross-linked peptides having four terminal residues compared to only two on a linear peptide, but this also limits the choice of protease as SCX is most effective when protonation sites are available at both the N- and C-termini. To avoid this limitation, we made use of successive enzymatic digests, initially a tryptic digest before enrichment of the cross-linked peptides followed by a second digest with Glu-C. Using this combination resulted in approximately 1.5-fold enrichment in the number of cross-linked peptides identified over trypsin alone. The exact figure was dependant on the sample analyzed [34]. An alternative to the use of SCX is the application of only selecting those ions with a charge >3+ for fragmentation. While this does not reduce the complexity of the sample, it has proved successful in increasing the number of cross-linked peptides detected as less time is spent on the acquisition of lower charged species which have a higher representation of linear peptides.

The modification of the amino acid side chains on reaction with cross-linking reagents is typically associated with an inability for enzymatic cleavage at this location. Trypsin, for example, is unable to carry out proteolysis after the amine of the lysine has been converted into an amide by bi-functional activited esters. This feature of cross-linked peptides makes it possible to use a non-specific digestion to enrich these peptides [39]. In a typical sample preparation, the addition of Proteinase K to a solution of cross-linked proteins will result in the complete digestion of the un-cross-linked regions of the sample. The disadvantage of this and related methods is that, as while this greatly aids the identification of cross-links against the background, the task of predicting cross-linked peptides sequences is greatly complicated by the lack of protease specificity. The implications of this will be discussed in more detail in Section 2.4.

2.2. Cross-linking reagents

The growth in the use of protein cross-linking coupled with mass spectrometry has led to an expansion in the number of cross-linking reagents in the literature with a variety of chemical specificity and distance between functional groups [40–46]. These have been designed to solve many of the different limitations that more traditional reagents have faced or to provide some kind of previously unmet utility. The diversity of the reagents now available is much larger than previously but there are still many limitations that continue to be the focus for development. The recent key developments lie in the types of chemistry available to undertake the cross-linking reaction and the novel designs that aid in the detection of the cross-linked peptides formed by the use of these reagents.

2.2.1. N-hydroxysuccinimide (NHS) activated ester

The use of N-hydroxysuccinimide (NHS) activated ester cross-linking reagents (see Fig. 4) continues to be a stable basis for a large number of cross-linking studies [47–50,34,51]. This is likely due to their commercial availability and the development of stable isotope labeled reagents to aid detection (see Section 2.4) [41]. NHS esters are primarily used for their ability to target the primary amines found in the side chain of lysine residues. The reactivity is aided by the position of the amine at the end of the side chain, reducing steric hindrance, while the polar nature of the residue means it is typically found on the surface of proteins or at interface sites. Typically NHS esters have been assumed to be specific in their reactivity, only targeting lysine residues, but there is increasing evidence that these reagents show a much broader range of targets with amino acids that have previously been ignored [52]. In a standard cross-linking reaction, the amount of the NHS ester and time for which it is incubated with the protein sample has to be optimized on an experiment-by-experiment basis as it varies with the lysine content and availability within the sample. To control the end point of the reaction, it is typically quenched with an amine containing buffer before the products are monitored by electrophoresis. Care must be taken to identify and control the formation of higher-order oligmers as these may lead to the detection of non-specific cross-linking products. If it is not possible to optimize conditions to avoid the formation of these products, it may be possible to remove them before analysis by additional purification using size exclusion chromatography under denaturing conditions [34].

A major drawback of these reagents is their ability to inhibit proteolysis by trypsin at the reaction site, resulting in larger peptides under these conditions. In proteins with a limited number of potential proteolysis sites, this is particularly problematic. This can be resolved by the use of alternative enzymes or a combination of enzymes during sample preparation (see Section 2.1.1).

2.2.2. Acid specific cross-linking reagents

The prevalence of reagents that focus on the coupling of lysine residues in mass spectrometry-coupled cross-linking experiments has resulted in increasing efforts to develop alternative linkers that target different residues [53]. The most significant recent development in cross-linking chemistry was the demonstration of two acid residue specific cross-linking reagents – adipic acid dihydrazide and pimelic acid dihydrazide (see Fig. 4) – which provide a promising addition to the toolkit of lysine (and cysteine) specific reagents that are currently available by overcoming the previous requirement of relatively low pH [54]. They also present an interesting potential starting point for the development of gas-phase cleavable versions (see Section 2.2.5) based on the same chemistry.

2.2.3. Arginine specific cross-linking reagents and protein–DNA interactions

An untapped potential of protein cross-linking reactions is the possibility of developing a mechanism for identifying protein–DNA or protein–RNA interactions. The use of diglyoxal compounds to investigate these interactions has been demonstrated in the analysis of the organization of the ribosome [55,56]. The application of these cross-linking reagents has been demonstrated for both the study of protein structure with the identification of arginine–arginine cross-links [57,58] in proteins and its use in the analysis of nucleic acid structure [59]. To aid the identification of

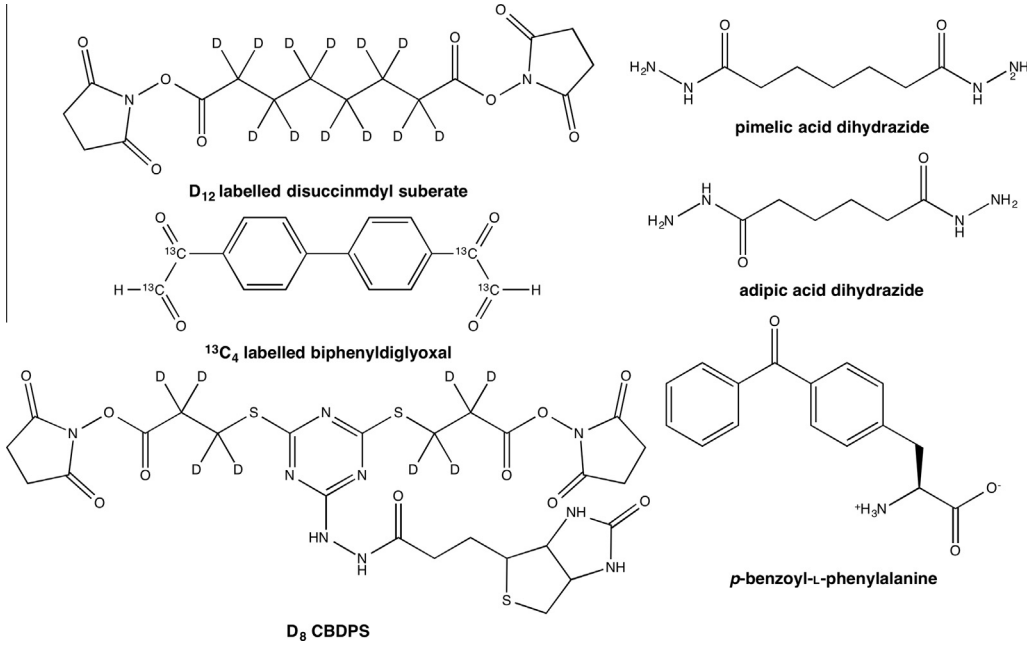

Fig. 4. Examples of the different cross-linking reagents discussed in this review. Disuccinimidyl suberate (DSS) is a lysine reactive cross-linking reagent, shown here in its deuterated form (see Section 2.4.2). Biphenyldiglyoxal is an arginine reactive cross-linking reagent, shown here with a carbon-13 label (see Section 2.2.3). CBDPS (Cyanurbiotindimercaptopropionylsuccinimide), brings together multiple different features to aid detection of cross-linked peptides by mass spectrometry (see Section 2.3.1). Adipic and pimelic acid dihydrazide are acid reactive cross-linking reagents; they are shown here as a pair to illustrate how cross-linking length can be varied (see Section 2.2.2). *p*-Benzoyl-L-phenylalanine is an example of a cross-linking reagent that can be genetically encoded into the protein (see Section 2.2.4).

cross-linked species, I developed a ^{13}C labeled form of biphenyldiglyoxal (see Fig. 4), from the Friedel–Crafts acylation reaction between biphenyl and $^{13}C_2$-acetyl chloride followed by oxidation of the terminal carbons using HBr/DMSO oxidation. Initial evaluation of diglyoxal-based cross-linkers for the identification of protein–protein interactions was confirmed by the incubation of the purified cross-linker with the target PyrR protein complex [60]. The range of molar ratios between the cross-linking reagentto protein were varied from 5:1 to 50:1 in 50 mM ammonium borate buffer pH 8.1. Monitoring the reaction by SDS–PAGE showed the clear formation of a cross-linked homotetramer in the case of PyrR. Using this reagent, it was possible to identify the formation of arginine-selective cross-links (see Fig. 5) in the human PyR complex [61,62] where previous attempts with NHS esters had been unsuccessful.

2.2.4. Genetically encoded cross-linking reagents

The development of genetically encoded *p*-benzoyl -L-phenylalanine (pBpa) [64] (see Fig. 4) using the amber codon (UAG) allows for the incorporation of a photo-activated cross-linking reagent into a specific gene within an organism. The ability to genetically encode a photo-reactive amino acid has several distinct advantages for the investigation of protein interactions and binding within the cell. The location of the amino acid can be used to directly probe where on the surface individual interactions happen. This is achieved by generating a series of mutants with the pBpa incorporated at different points along the amino acid sequence. Only when the activated amino acid can be cross-linked on the surface of the protein and in contact with a second species will that interaction be detected. By studying the results of these in turn, we can then identify the environment surrounding the protein. This method greatly reduces the complexity of the problem in identifying the peptides, by limiting the cross-links to only one possible position from the mutant protein, therefore the identification of the proteolytic product by mass spectrometry is not significantly more difficult than a typical linear peptide search. The application of this method has been aided by the development of an isotope labeled form of the amino acids [65], meaning that only the cross-linked peptides and the peptides within the site of incorporation will contain the isotope label (see Fig. 6).

These benefits were utilized in the development of Hekate's [63] 'amber codon mode'. Once enabled, the software will request details of known peptides containing the amber codon residue alongside information on the type of isotope label used and the mass of the amber codon residue both before and after cross-linking (available from http://github.com/MRC-LMB-MassSpec/Hekate at publication). The peptide containing the amber codon can be predicted from the amino acid sequence of the bait protein and the specificity of the protease used. Using this method, it was possible to identify the site of cross-linking between Sdo1 – a yeast authologue of SBDS, a protein implicated in Shwachman–Bodian–Diamond syndrome – and Efl1 (see Fig. 7).

2.2.5. Cleavable cross-linking reagents

A variety of cleavable cross-linking reagents have been developed to aid identification and sequencing of cross-linked peptides. Some of the oldest forms of these reagents make use of a reducible di-sulfide bridge within the molecule [42]. The protein is cross-linked in the usual way to stabilize non-covalent interactions and then digested; and the sample is then analyzed by mass spectrometry. The convenience of this method, though, is that cross-linked peptides can be re-analyzed following reduction with a thiol compound. The products of this reaction still have a detectable modification, but now appear as linear peptides.

Using differential peptide mapping (comparing the sample spectra before and after reduction and identifying peaks that are different between conditions), it is possible to characterize the products of cross-linking. Cross-linked peptides often fragment poorly and the spectra are much more difficult to analyze than that of a linear peptide. Difficulties arise as the two peptides chains fragment independently often with one dominating, e.g. from the

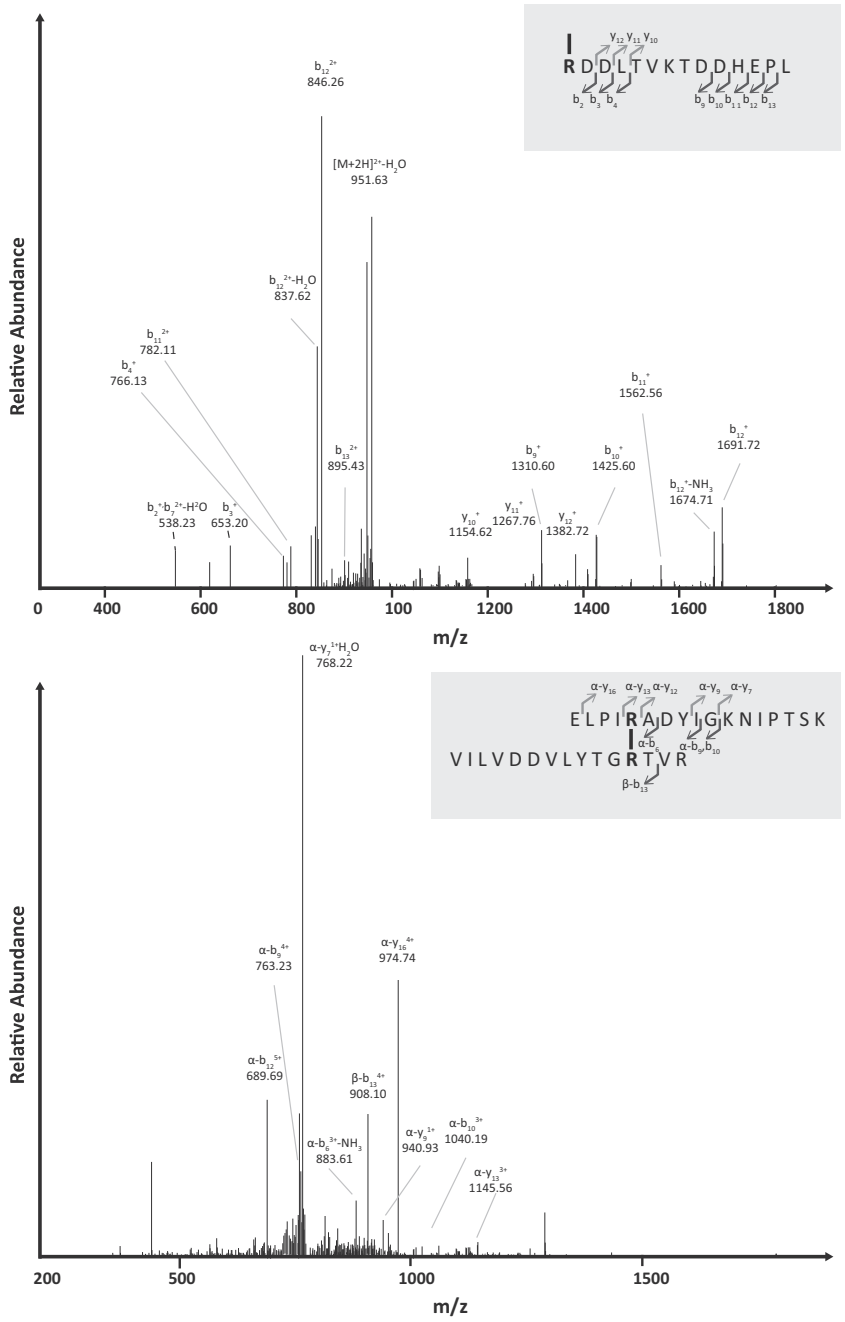

Fig. 5. (Top) Product ion spectra of the RDDLTVKTDDHEPL monolinked peptide, confirming the incorporation of the BPG monolink. (Bottom) Product ion spectra of the detected cross-linked peptide. The precursor ion ELPIRADYIGKNIPTSK-VILVDDVLYTGRTVR was detected with a mass accuracy of 0.6 ppm. These peptides were identified using Hekate [63].

N-terminal side of proline, thus hindering the confirmation of the second peptide sequence. The benefit of the use of differential peptides analysis, therefore, is that it allows for the resultant linear peptides to be fragmented with greater ease when compared to that of the precursor cross-linked peptide. Cross-linking reagents that cleave under UV-light also exist and provide a similar utility [66].

A more recent application of cleavable cross-linkers is the development of cross-linking reagents that can be cleaved within the mass spectrometer [67]. These cross-linking reagents are fragmented into two separate peptide chains by collision-induced dissociation (CID) within the mass spectrometer. The linear peptide products produced in the gas phase can then be selected and isolated before a second round of fragmentation. This fragmentation then provides information on the amino acid sequence for each constituent linear peptide. This process is a compromise: the spectra produced are more easily assigned, but it results in a loss in signal intensity and increased acquisition time. The data-dependent acquisition techniques needed to undertake this kind of experiment are common to a range of mass spectrometers. These fragmentation patterns are then suitable for identification by the use of standard database solutions found in proteomic software.

2.3. Cross-linking reagents and sample enrichment

As already described, the use of chromatography to enrich for cross-linked peptides (see Sections 2.1.2 and 2.1.3) has been shown to be effective, but due to the similarities of the linear and

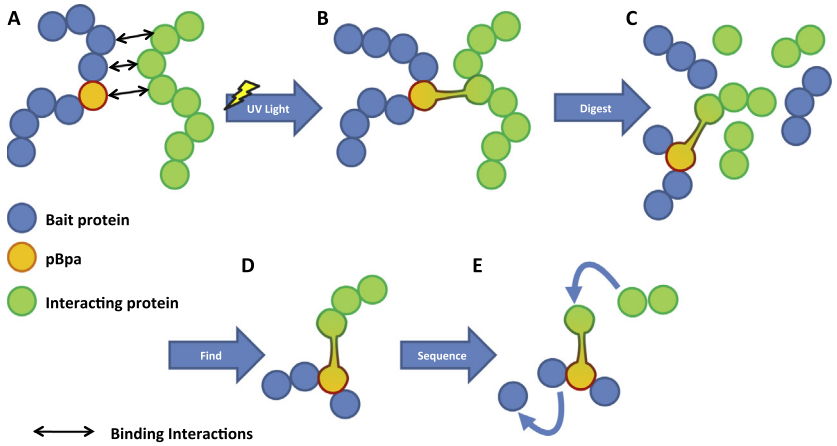

Fig. 6. The analysis of protein interactions using genetically encoded cross-linking reagents uses many of the same steps as a standard cross-linking strategy (see Fig. 1). The significant difference is the cross-link will only be initiated from the position where the photo-reactive amino acid (pBpa) has been included. This reduces the complexity of the final analysis.

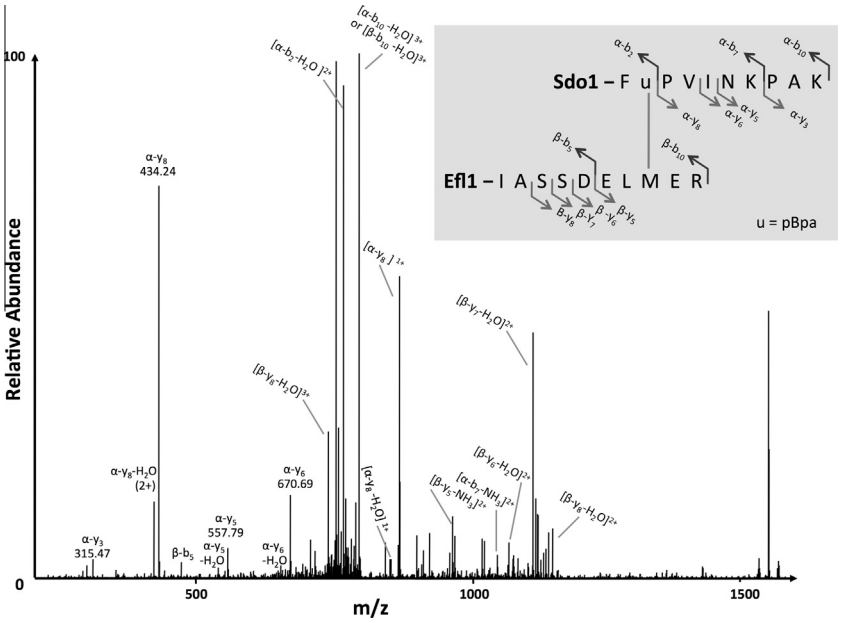

Fig. 7. The product ion spectrum of the isolated peptide as identified by Hekate and generated by the use of an amber codon mutant to incorporate pBpa into a bait protein (Sdo1). Amino acid *u* = pBpa.

cross-linked peptides they do not provide an exclusive method to separate one from the other. The problem is further complicated as the linear peptides are very much in excess of the target cross-linked peptides. The development of affinity-based purification with the potential to exclusively enrich for peptides containing the cross-linking reagent have the potential to provide dramatic benefits to the study of protein structure.

2.3.1. Biotin labeled cross-linking reagents

The development of PIR [40] and CBDPS [68] (see Fig. 4) cross-linking reagents brought together several different strategies into a single reagent. In particular the inclusion of a biotin tag, as previously demonstrated by Alley et al. [69], within the bridge of the cross-linking reagent allows for the cross-linked peptides to be purified with a much higher specificity than is possible by other techniques. This tag is combined with the inclusion of three sites that can be simultaneously cleaved within the gas-phase by low-energy collision induced dissociation. This releases a reporter ion, and the separated chains of the cross-linked peptides for sequencing. A convenient result of this is that linear peptides, loop-link peptides and cross-linked peptides will give very different products as a result of this first fragmentation step, while the reporter ion aids in providing a key marker for informatics software to identify relevant spectra. The product ions that represent the individual peptides chains can undergo another round of fragmentation to provide sequence information, as previously discussed in Section 2.2.5.

2.3.2. Click-based affinity tags

An alternative to the incorporation of a biotin tag has been demonstrated by the incorporation of an azide functional group within the linking reagent [44]. The inclusion of an azide is convenient as it is both much smaller than a biotin tag and enables for the use of a wide range of ‘click chemistry’ to capture the

cross-linked peptides. In the initial study, an azide reactive cyclooctyne resin (ARCO-resin) was used to enrich for cross-linked peptides. The resin can then be washed to remove any non-specific binding from linear peptides or other contaminants. The cross-linked peptides can then be released from the resin by the reduction of a disulfide bond in the linker bound to the resin [70].

2.4. Detection

2.4.1. Software

There has been a long history in the use of peptide databases to analyze proteomic data [28]. It was therefore inevitable that attempts would be made to adapt these techniques to the development of software packages to provide the same functionality for the analysis for cross-linked proteome data.

Cross-linked peptides pose a unique set of problems to the application of database searches. The size of the problem (see Section 2) is significantly larger and scales much more poorly than that for linear peptides. The peptides fragment in a different manner to linear peptides and the number of peptides that represent a particular interaction are non-stoichiometric, thus reducing the statistical power of any result from a cross-linking experiment. The development of reagents with multiple MS-directed features – for example PIR [40] and CBDPS [68] reagents (see Section 2.3.1) include reporter ions and isotope labels – provide potential improvements in detection. The impact of this wide variety of cross-linking methods means that the more complex cross-linking reagents need quite specific software e.g. BLinks and the identification of PIRs [71] or DXMSMS [72] for the analysis of isotope labeled CID cleavable cross-linked products.

Several attempts have been made to simplify the identification of the cross-linked peptides including fast-sorting [35], xQuest's [73] 'ion-tag' method and my own Hekate [63], which makes use of indexed databases.

The 'ion-tag' method employed by xQuest searches each spectra against a database of linear peptides in a precursor mass independent manner. The results of this initial search then identify which potential linear peptides can be combined to form potential cross-linked peptides. This greatly reduces the complexity of the initial database needed to search the cross-linked proteomic data.

In the development of Hekate [63], a different approach was taken. Hekate interfaces to a SQL database by the use of the Perl DBI modules. The database employed provides several indexing methods to aid the rapid searching of data. Futhermore, many databases have already been developed with a significant focus on speed and large data sets, thus making them potentially suitable for the exhaustive search of all potential cross-linked peptides. Conveniently, the DBI in Perl provides a unified interface to facilitate changes in the database used. During the initial stages of development, SQLite was used for convenience; switching from SQLite to MySQL resulted in a reduction of 26% in the time to reanalyze the data from the original publication [63] with no further optimization. While it has not yet been realized, this demonstrates the potential of undertaking this on much larger datasets by the application of the appropriate database choice and supporting hardware. Moreover, the use of a client/server based database such as MySQL opens up the possibility of using high-performance clusters in the analysis of data.

2.4.2. Isotope labeling

Identification of cross-linked peptides computationally by the scoring of each individual spectra is time consuming and wasteful of computer resources when we consider that, without highly specific enrichment, the majority of species detected will reflect the linear peptides. Therefore, techniques that aid in the detection of cross-linked peptides from the precursor ion mass are particularly valuable.

A convenient method to achieve this result is the use of oyxgen-18 labeling [74]. The proteolytic digestion of a sample is undertaken in oxygen-18-enriched buffer ($H_2^{18}O$) with trypsin will incorporate two oxygen-18 atoms in the C-terminus of the peptides. As cross-linked peptides have two C-termini, this results in the inclusion of four oxygen-18 atoms. The cross-linked peptides can then be identified either by a characteristic mass shift or by the 1:1 mixture of labeled and unlabeled samples. This results in characteristic doublets in the spectra with an 8 Da spacing, whereas linear peptides will only show a 4 Da doublet. The benefits of this method are that it is fast and can easily be adapted to a wide variety of cross-linking reagents. Software packages can then be designed to recognize these doublets and therefore quickly identify linear and cross-linked ions from the precursor spectra. Data-dependent acquisition also provides the opportunity to selectively target only those ions that show an appropriate doublet for fragmentation, potentially providing the opportunity to acquire less linear peptide data and more data specific to protein interactions. One issue that can complicate the analysis is incomplete incorporation of two oxygen-18 atoms into each carboxylate of the C-terminus. This can occur since the product is required to rebind to the protease in order to facilitate the incorporation of the second oxygen-18 atom.

A popular alternative strategy to the addition of oxygen-18 labels during proteolytic digestion is the application of isotope labeled cross-linking reagents [41,75]. By synthesizing two versions of a cross-linker, one with and one without the incorporation of a stable isotope, it is possible to use the two in a known ratio to provide a doublet on analysis by mass spectrometry. Unlike oxygen-18 labeling, the use of an isotope label within the cross-linker also aids the detection of monolinks and loop-links, not just cross-linked peptides with two C-termini. This may or may not be of benefit depending on the research undertaken. The placement of the label within the cross-linking regent also provides the opportunity to use a wider variety of stable isotopes to label samples. In particular, the use of deuterium as a stable isotope label has been used in a wide range of studies due to the commercial availability of bi-functional NHS-activated esters [7,19,34,51,76,73,75,77]. While it has previously been noted that deuterium labeling can cause a detectable difference in retention time when separating peptides by reversed-phase liquid chromatography because of the chromatographic isotope effect [78], this has not been reported to have had a significant impact when identifying cross-linked peptides by this method.

The formation of a 'mass doublet' by the use of isotopic labels has been a particularly valuable development as it provides a visual cue for the identification of cross-linked peptides, a trigger for data-dependent acquisition, and a filter for the analysis that is easily automated in software. The convenience of this method has led to the wide variety of software written to undertake the analysis of cross-linked protein data applying these methods, supporting some form of isotope labeling in an initial filtering of the data before scoring. This results in both faster processing of data produced by protein cross-linking and higher confidence in the identification of cross-links due to both the labeled and unlabeled peptide spectra providing complementary information. Both xQuest and Hekate use the shift of peptide fragments between the heavy and light forms of cross-linked peptides to identify when a particular fragment contains the cross-link and to reduce the complexity of the spectra. As cross-links are often low in abundance, this cross-validation approach is helpful for hard-to-assign fragmentation spectra.

The success of stable isotope labeling in XL-MS has led to the prevalence of one-to-one ratios in experimentation as these result

in a mass doublet with the same ratio of intensity. More recently, though, it has been shown that this does not result in the highest number of identified peptides [79]. In an analysis by Fischer et al., they undertook the comparison of a 1:1, a 1:2 and a 1:4 ratio between the two forms of the cross-linker. The results demonstrated that the highest number of cross-linked peptides were identified with a ratio of 1:4. This is presumably because it provided the highest intensity signal in terms of the 'light' peptides, thus aiding identification. How dependent this result was on the computational methods they employed, however, is not clear.

2.5. False discovery rates (FDR)

The challenge in gaining reliable insight from the digestion of cross-linked protein samples is not just one of identification. Once a potential cross-linked peptide has been detected and its sequence has been established, the next step is validation of the result. This will typically be achieved by undertaking further research using a complementary technique, but if we are to use XL-MS on large and complex systems then this may be impractical. The application of target-decoy strategies as implemented by xProphet [80] (alongside xQuest), Plink [81] and later Hekate allow for the provision of false discovery rates (FDR). To calculate the FDR, two databases of peptides are generated: the target database of known sequences, and a second decoy database. The detected peptides are then scored against potential matches in both databases (and sometimes a hybrid database containing cross-links). The FDR is established by counting the number of spectra that matched the decoy database with a score equal to or greater than the spectra of interest. This is then expressed as a percentage of the total number of spectra that matched either database at that score or above. In a typical study, any result with an FDR <1% is considered of interest. The addition of false discovery rates to the analysis of cross-linking methods is a vital tool for the comparision of samples and for the quantitative measurement of the quality of data produced.

2.6. Quantitative cross-linking

Historically, the use of stable isotopes was undertaken to aid in the identification of cross-linked peptides (see Section 2.4.2); however, it has recently been demonstrated that the same stable isotope cross-linking reagents can be used to provide relative quantification of interactions between samples. If the two protein samples in different conditions are reacted separately, the process can be used to label the cross-linked products depending on the conditions in which they were generated [79]. Comparison of the abundance of cross-linked peptides provides quantitative information on the organization of the protein population. Because of the immaturity of the technique, there is currently only a limited amount of support in terms of software for the application of this method; however, the provision of XiQ [79] with the original publication provides a minimal framework to further establish and apply the technique [12]. This limited integration of support in cross-linking currently means that quantification is undertaken manually [12].

3. Conclusions

Mass spectrometry coupled with protein cross-linking has and continues to demonstrate considerable value in the characterization of the interactions between proteins. Even the most basic application to the stabilization of transient interactions within protein complexes becomes immensely powerful when applied to the investigation of novel binding partners that would otherwise not be detected. At the same time, some parts of the field are developing at a rapid pace.

The wide range of applications of XL-MS continues to grow. The rise of systems biology has led to a fundamental need to analyze more complex systems more rapidly. Even relatively immature methods have had a large impact on what is achievable in the field. This advancement is demonstrated by the application of quantitative cross-linking: despite there being no integrated solutions currently available, their application to the analysis of the organization of an F-type ATPase has provided insight into a highly complex problem [82]. As more integrated methods for the analysis of these data are developed, we can expect further application and scope of XL-MS to the analysis of these systems.

The availability of methods [73,83] to undertake the large-scale analysis of cross-linked peptides from complex mixtures has been a major development for the field. This has led to the ability to build protein interaction networks [84] from *in vivo* cross-linking experiments. The value of these techniques is demonstrated by the current efforts already committed to build a proteome scale static map of interactome using Y2H assays [85,86] and the analysis of sequence co-evolution [87]. XL-MS could achieve the same results in a context specific manner providing an *in vivo* insight that builds upon the current static interactome data to include dynamic information of how interactions between proteins change in response to a specific stimuli or perturbation. There is no doubt that the biggest challenges for this kind of study is expanding the data set from around 4300 proteins within *Escherichia coli* proteome to nearly 90,000 within the human proteome [35]. Despite this, high performance computing remains an untapped possibility for this research field and, if successfully applied, has the potential to revolutionize the way we analyze protein networks and protein complex organization to provide a novel systems approach to structural biology.

Acknowledgements

ANH would like think T. Perica and S.A. Teichmann for the donation of the PyrR protein, and N. Basse and A. Warren for the donation of cross-linked Sdo1–Efl1 complex. This work was support by the Medical Research Council – United Kingdom.

References

[1] C.V. Robinson, E.W. Chung, B.B. Kragelund, J. Knudsen, R.T. Aplin, F.M. Poulsen, C.M. Dobson, J. Am. Chem. Soc. 118 (1996) 8646–8653.
[2] D. Suckau, M. Mak, M. Przybylski, Proc. Natl. Acad. Sci. 89 (1992) 5630–5634.
[3] G. Xu, M.R. Chance, Chem. Rev. 107 (2007) 3514–3543.
[4] M.J. Chalmers, S.A. Busby, B.D. Pascal, Y. He, C.L. Hendrickson, A.G. Marshall, P.R. Griffin, Anal. Chem. 78 (2006) 1005–1014.
[5] Z. Zhang, D.L. Smith, Protein Sci. 2 (1993) 522–531.
[6] F. Stengel, R. Aebersold, C.V. Robinson, Mol. Cell. Proteomics 11 (2012) R111–014027.
[7] Z.A. Chen, A. Jawhari, L. Fischer, C. Buchen, S. Tahir, T. Kamenski, M. Rasmussen, L. Lariviere, J.-C. Bukowski-Wills, M. Nilges, et al., EMBO J. 29 (2010) 717–726.
[8] A. Sinz, Mass Spectrom. Rev. 25 (2006) 663–682.
[9] A. Sinz, Exp. Rev. Proteomics 11 (2014) 733–743.
[10] J. Rappsilber, J. Struct. Biol. 173 (2011) 530–540.
[11] M. Sharon, A. Sinz, Anal. Biomol. Interact. Mass Spectrom. (2015) 55–79.
[12] C. Schmidt, C.V. Robinson, Nat. Protoc. 9 (2014) 2224–2236.
[13] A. Leitner, T. Walzthoeni, R. Aebersold, Nat. Protoc. 9 (2014) 120–137.
[14] H. Zahn, Angew. Chem. 67 (1955) 561–572.
[15] H.C. Berg, J.M. Diamond, P.S. Marfey, Science 150 (1965) 64–67.
[16] G.E. Davies, G.R. Stark, Proc. Natl. Acad. Sci. 66 (1970) 651–656.
[17] G. Fleet, R. Porter, J. Knowles, Nature 224 (1969) 511–512.
[18] G. Fleet, J. Knowles, R. Porter, Biochem. J. 128 (1972) 499–508.
[19] S. Mädler, C. Bich, D. Touboul, R. Zenobi, J. Mass Spectrom. 44 (2009) 694–706.
[20] A.J. Lomant, G. Fairbanks, J. Mol. Biol. 104 (1976) 243–261.
[21] L. Zhang, S. Rayner, N. Katoku-Kikyo, L. Romanova, N. Kikyo, Biochem. Biophys. Res. Commun. 361 (2007) 611–614.
[22] C. Adrain, M. Zettl, Y. Christova, N. Taylor, M. Freeman, Science 335 (2012) 225–228.
[23] R. Aebersold, M. Mann, Nature 422 (2003) 198–207.
[24] B.F. Cravatt, G.M. Simon, J.R. Yates Iii, Nature 450 (2007) 991–1000.

[25] H. Mohammed, C. DSantos, A.A. Serandour, H.R. Ali, G.D. Brown, A. Atkins, O.M. Rueda, K.A. Holmes, V. Theodorou, J.L. Robinson, et al., Cell Rep. 3 (2013) 342–349.
[26] J. Déjardin, R.E. Kingston, Cell 136 (2009) 175–186.
[27] F. Pourfarzad, A. Aghajanirefah, E. de Boer, S. Ten Have, T. Bryn van Dijk, S. Kheradmandkia, R. Stadhouders, S. Thongjuea, E. Soler, N. Gillemans, et al., Cell Rep. 4 (2013) 589–600.
[28] D.J. Pappin, P. Hojrup, A.J. Bleasby, Curr. Biol. 3 (1993) 327–332.
[29] J.K. Eng, A.L. McCormack, J.R. Yates, J. Am. Soc. Mass Spectrom. 5 (1994) 976–989.
[30] J.S. Cottrell, U. London, Electrophoresis 20 (1999) 3551–3567.
[31] C.C. Wu, M.J. MacCoss, Curr. Opin. Mol. Ther. 4 (2002) 242–250.
[32] H. Choi, A.I. Nesvizhskii, J. Proteome Res. 7 (2007) 47–50.
[33] M.M. Young, N. Tang, J.C. Hempel, C.M. Oshiro, E.W. Taylor, I.D. Kuntz, B.W. Gibson, G. Dollinger, Proc. Natl. Acad. Sci. 97 (2000) 5802–5806.
[34] A. Toste Rêgo, A.N. Holding, H. Kent, M.H. Lamers, EMBO J. 32 (2013) 1334–1343.
[35] E.V. Petrotchenko, C.H. Borchers, Proteomics 14 (2014) 1987–1989.
[36] J.V. Olsen, S.-E. Ong, M. Mann, Mol. Cell. Proteomics 3 (2004) 608–614.
[37] A. Leitner, R. Reischl, T. Walzthoeni, F. Herzog, S. Bohn, F. Förster, R. Aebersold, Mol. Cell. Proteomics 11 (2012). M111–014126.
[38] R. Fritzsche, C.H. Ihling, M. Götze, A. Sinz, Rapid Commun. Mass Spectrom. 26 (2012) 653–658.
[39] E.V. Petrotchenko, J.J. Serpa, D.B. Hardie, M. Berjanskii, B.P. Suriyamongkol, D.S. Wishart, C.H. Borchers, Mol. Cell. Proteomics 11 (2012). M111–013524.
[40] X. Tang, J.E. Bruce, Mol. Biosyst. 6 (2010) 939–947.
[41] D. Müller, P. Schindler, H. Towbin, U. Wirth, H. Voshol, S. Hoving, M. Steinmetz, Anal. Chem. 73 (2001) 1927–1934.
[42] K.L. Bennett, M. Kussmann, M. Mikkelsen, P. Roepstorff, P. Bjrk, M. Godzwon, P. Srensen, Protein Sci. 9 (2000) 1503–1518.
[43] B.W. Sutherland, J. Toews, J. Kast, J. Mass Spectrom. 43 (2008) 699–715.
[44] H. Buncherd, M.A. Nessen, N. Nouse, S.K. Stelder, W. Roseboom, H.L. Dekker, J.C. Arents, L.E. Smeenk, M.J. Wanner, J.H. van Maarseveen, et al., J. Proteomics 75 (2012) 2205–2215.
[45] L. Yang, X. Tang, C.R. Weisbrod, G.R. Munske, J.K. Eng, P.D. von Haller, N.K. Kaiser, J.E. Bruce, Anal. Chem. 82 (2010) 3556–3566.
[46] N. Hino, Y. Okazaki, T. Kobayashi, A. Hayashi, K. Sakamoto, S. Yokoyama, Nat. Methods 2 (2005) 201–206.
[47] I. Karadzic, J. Maupin-Furlow, M. Humbard, L. Prunetti, P. Singh, D.R. Goodlett, Proteomics 12 (2012) 1806–1814.
[48] D.M. Schulz, S. Kalkhof, A. Schmidt, C. Ihling, C. Stingl, K. Mechtler, O. Zschörnig, A. Sinz, Funct. Bioinf. 69 (2007) 254–269.
[49] J. Pettelkau, I. Thondorf, S. Theisgen, H. Lilie, T. Schröder, C. Arlt, C.H. Ihling, A. Sinz, J. Am. Soc. Mass Spectrom. 24 (2013) 1969–1979.
[50] M. Götze, J. Pettelkau, S. Schaks, K. Bosse, C.H. Ihling, F. Krauth, R. Fritzsche, U. Kühn, A. Sinz, J. Am. Soc. Mass Spectrom. 23 (2012) 76–87.
[51] Y. Liu, H.K. Salter, A.N. Holding, C.M. Johnson, E. Stephens, P.J. Lukavsky, J. Walshaw, S.L. Bullock, Genes Dev. 27 (2013) 1233–1246.
[52] S. Kalkhof, A. Sinz, Anal. Bioanal. Chem. 392 (2008) 305–312.
[53] P. Novak, G.H. Kruppa, Eur. J. Mass Spectrom. 14 (2008) 355.
[54] A. Leitner, L.A. Joachimiak, P. Unverdorben, T. Walzthoeni, J. Frydman, F. Frster, R. Aebersold, Proc. Natl. Acad. Sci. 111 (2014) 9455–9460.
[55] H.F. Noller, J.B. Chaires, Proc. Natl. Acad. Sci. 69 (1972) 3115–3118.
[56] H.F. Noller, Biochemistry 13 (1974) 4694–4703.
[57] M.O. Lederer, R.G. Klaiber, Bioorg. Med. Chem. 7 (1999) 2499–2507.
[58] Q. Zhang, E. Crosland, D. Fabris, Anal. Chim. Acta 627 (2008) 117–128.
[59] Q. Zhang, E.T. Yu, K.A. Kellersberger, E. Crosland, D. Fabris, J. Am. Soc. Mass Spectrom. 17 (2006) 1570–1581.
[60] C.L. Turnbough, R.L. Switzer, Microbiol. Mol. Biol. Rev. 72 (2008) 266–300.
[61] A.N. Holding, E. Stephens, in: 59th ASMS Conference Proceedings, 2011.
[62] S.L. Teichman, A. Ferrick, S.G. Kim, J.A. Matos, L.E. Waspe, J.D. Fisher, J. Am. Coll. Cardiol. 10 (1987) 633–641.
[63] A.N. Holding, M.H. Lamers, E. Stephens, J.M. Skehel, J. Proteome Res. 12 (2013) 5923–5933.
[64] I.S. Farrell, R. Toroney, J.L. Hazen, R.A. Mehl, J.W. Chin, Nat. Methods 2 (2005) 377–384.
[65] T. AshtonáCropp et al., Mol. BioSyst. 4 (2008) 934–936.
[66] L. Yang, X. Tang, C.R. Weisbrod, G.R. Munske, J.K. Eng, P.D. von Haller, N.K. Kaiser, J.E. Bruce, Anal. Chem. 82 (2010) 3556–3566.
[67] E.J. Soderblom, M.B. Goshe, Anal. Chem. 78 (2006) 8059–8068.
[68] E.V. Petrotchenko, J.J. Serpa, C.H. Borchers, Mol. Cell. Proteomics 10 (2011). M110–001420.
[69] S.C. Alley, F.T. Ishmael, A.D. Jones, S.J. Benkovic, J. Am. Chem. Soc. 122 (2000) 6126–6127.
[70] M.A. Nessen, G. Kramer, J. Back, J.M. Baskin, L.E. Smeenk, L.J. de Koning, J.H. van Maarseveen, L. de Jong, C.R. Bertozzi, H. Hiemstra, et al., J. Proteome Res. 8 (2009) 3702–3711.
[71] M.R. Hoopmann, C.R. Weisbrod, J.E. Bruce, J. Proteome Res. 9 (2010) 6323–6333.
[72] E.V. Petrotchenko, K.A. Makepeace, C.H. Borchers, Curr. Protoc. Bioinform. 48 (2014) 8.18.1–8.18.19.
[73] O. Rinner, J. Seebacher, T. Walzthoeni, L. Mueller, M. Beck, A. Schmidt, M. Mueller, R. Aebersold, Nat. Methods 5 (2008) 315–318.
[74] J.W. Back, V. Notenboom, L.J. de Koning, A.O. Muijsers, T.K. Sixma, C.G. de Koster, L. de Jong, Anal. Chem. 74 (2002) 4417–4422.
[75] K.M. Pearson, L.K. Pannell, H.M. Fales, Rapid Commun. Mass Spectrom. 16 (2002) 149–159.
[76] J. Lee, S. Ding, T.B. Walpole, A.N. Holding, M.G. Montgomery, I.M. Fearnley, J.E. Walker, J. Biol. Chem. (2015). jbc–M115.
[77] B.X. Huang, H.-Y. Kim, C. Dass, J. Am. Soc. Mass Spectrom. 15 (2004) 1237–1247.
[78] R. Zhang, C.S. Sioma, R.A. Thompson, L. Xiong, F.E. Regnier, Anal. Chem. 74 (2002) 3662–3669.
[79] L. Fischer, Z.A. Chen, J. Rappsilber, J. Proteomics 88 (2013) 120–128.
[80] T. Walzthoeni, M. Claassen, A. Leitner, F. Herzog, S. Bohn, F. Förster, M. Beck, R. Aebersold, Nat. Methods 9 (2012) 901–903.
[81] B. Yang, Y.-J. Wu, M. Zhu, S.-B. Fan, J. Lin, K. Zhang, S. Li, H. Chi, Y.-X. Li, H.-F. Chen, et al., Nat. Methods 9 (2012) 904–906.
[82] C. Schmidt, M. Zhou, H. Marriott, N. Morgner, A. Politis, C.V. Robinson, Nat. Commun. 4 (2013).
[83] H. Zhang, X. Tang, G.R. Munske, N. Tolic, G.A. Anderson, J.E. Bruce, Mol. Cell. Proteomics 8 (2009) 409–420.
[84] C.R. Weisbrod, J.D. Chavez, J.K. Eng, L. Yang, C. Zheng, J.E. Bruce, J. Proteome Res. 12 (2013) 1569–1579.
[85] T. Rolland, M. Taşan, B. Charloteaux, S.J. Pevzner, Q. Zhong, N. Sahni, S. Yi, I. Lemmens, C. Fontanillo, R. Mosca, et al., Cell 159 (2014) 1212–1226.
[86] L. Bonetta, Nature 468 (2010) 851–854.
[87] T.A. Hopf, C.P.I. Schärfe, J.P.G.L.M. Rodrigues, A.G. Green, O. Kohlbacher, C. Sander, A.M.J.J. Bonvin, D.S. Marks, eLife 3 (2014).